Scribe Notes for 4/8/03

HMMs and Gene Finding

Biological Problem – Predict genes by using tools to search the genome

I Bacteria

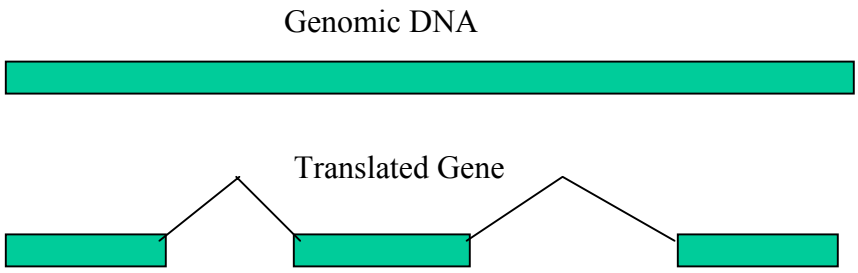
- genomes are small and compact
- gene is usually single unit starting with ATG and ending with stop codon
- if ATG and STOP codons in frame is your only criteria you will get many false positives.
    - will get nested genes, (small ORFs inside a larger ORF)
    - pseudogenes and overlapping ORFs
- early programs simply found all ORFs with ATG and STOP and took the largest possible ORFs in the case of nesting and overlapping'
- this was fairly successful, predicted 6200 genes in yeast and >5000 are real

- GLIMMER
    - program derived from TIGR
    - tries to access whether codon pattern in ORF is consistent with known genes
    - uses 5^{th} order HMM, likelihood of next base dependent on previous 5 positions, thus it looks at 2 codons at a time (actually uses variable order HMM depending on the situation)

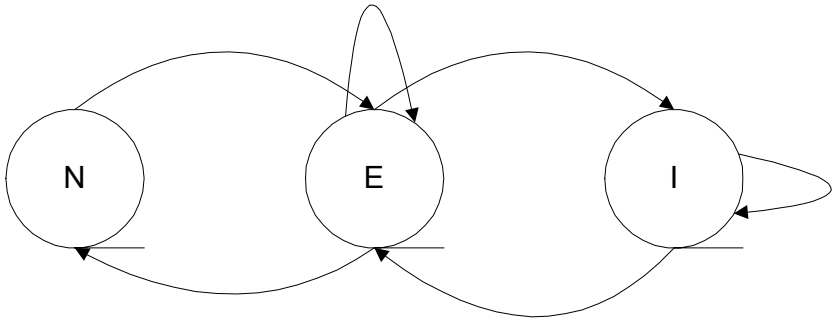
- Annotating the bacterial genome
    - Run GLIMMER (train from homologs)
    - BLAST predicted genes (if homologs, more likely to be real)

- BLAST rest of genome to see if missed anything likely to be a gene
- some programs look for initiation sites before ATG although they haven't led to better results

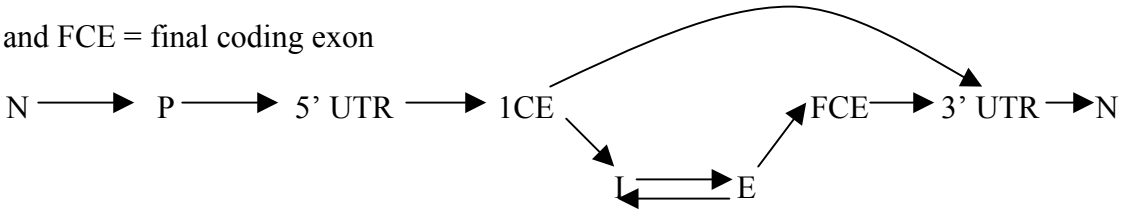
- Higher Organisms
  - much more difficult in eukaryotic cells, main problem is splicing.
  -



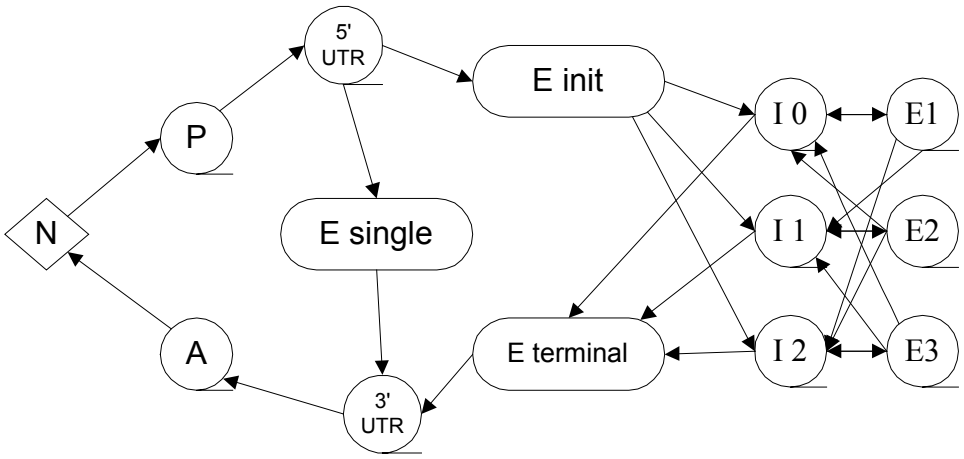
  - Large part of genome within gene is not actually translated with gene, this makes predicting much harder
  - Simple model: N = non coding region, E = exon, I = intron



  - More complex model, adds a few features, adds P = promoter, 5' UTR = 5' untranslated region, 1CE = first coding exon, 3'UTR = 3' untranslated region, and FCE = final coding exon
  -



- The above model is implemented because of differences in internal exons and the exons on the ends, namely the location of splice sites (1CE has 3' site, FCE has 5' site, and E has sites on both sides)
- Actually, have to consider differences in introns and exons depending on frame, as well as differences in single exon vs. many exons. This leads to the widely used model GenScan model



- This model is much better than earlier model, it more accurately describes splicing and leads to better results.
- For distantly related organisms, exons are much better conserved than introns, the above model can be modified to take this into account, leads to better results.

- Genomic Browser
  - A tool to browse through a genome and see various features.
  - Two widely used ones are GenScan and SLAM
  - They show that predicted gene from various programs and cDNA data so that the user can see all the information about a region and make an educated decision about the likelihood of a gene.

- What is a good prediction ???
    - This is a widely debated topic.
    - It is best to predict the real gene, but this is very hard to do
    - Is it better to get correct number of exons, or more accurate exons
    - Currently, prediction is not good enough to believe without testing. Thus, a prediction program that allows easy in vivo testing is the best.